Comparaison de mesures d'intérêt pour l'alignement de hiérarchies textuelles

Jerome David, Fabrice Guillet, Henri Briand

Comparaison de mesures d'intérêt pour l'alignement de hiérarchies textuelles

Jérôme David[1], Fabrice Guillet[1], Henri Briand[1]

Ecole Polytechnique de l'université de Nantes, LINA FRE CNRS 2729
Site de la Chantrerie, 44306 Nantes cedex 3
jerome.david-at-univ-nantes.fr

Résumé : Nous avons proposé une approche asymétrique et extensionnelle d'alignement qui permet d'extraire des relations (équivalences mais aussi subsomptions) entre les concepts de deux taxonomies textuelles. Cette approche repose sur l'idée qu'un concept A sera plus spécifique ou équivalent à un autre concept B si le vocabulaire utilisé dans les documents de A a tendance à être inclus dans celui de B. Dans le but d'évaluer de telles tendances, notre approche s'appuie sur le paradigme des règles d'association, et bénéficie des mesures d'intérêts développées dans ce contexte. Ainsi, nous proposons dans ce papier une comparaison expérimentale des alignements obtenus par différentes mesures d'intérêt. Ces mesures ont été sélectionnées selon des critères concernant leurs propriétés et leur sémantique. La première expérimentation concerne l'évaluation des résultats obtenus selon les différentes mesures sélectionnées. La deuxième s'intéresse aux distributions des valeurs obtenues par les mesures sur deux alignements : un contenant que des relations pertinentes et un autre consitué de relations non-pertinentes. Les résultats montrent que l'intensité d'implication obtient les meilleurs résultats.
Mots-clés : Alignement de taxonomies textuelles, règles d'association, mesures d'intérêt

1 Introduction

Sur le Web, les ressources, souvent sous forme textuelles, tendent à être organisées de manière hiérarchique. Des exemples de ce type de structuration sont les annuaires Web (Yahoo.com, OpenDirectory) et aussi les catalogues des boutiques en ligne (Amazon.com, Alapage.fr etc.). De plus, avec l'avènement du Web sémantique, les structures hiérarchiques sont aussi représentées au travers des ontologies OWL.

Même si l'organisation hiérarchique des contenus aide à structurer l'information et les connaissances disponibles, le Web demeure très hétérogène et dispersé. Les échanges de données ainsi que la communication entre les programmes ou les agents logiciels utilisant des données hiérarchisées reste ainsi très difficile. Dans le but de résoudre un tel problème d'interopérabilité, on doit être capable de comparer ces hiérarchies. Ainsi, de nombreuses méthodes d'alignement ont été proposées dans la littérature (Kalfoglou & Schorlemmer, 2003), (Rahm & Bernstein, 2001), (Shvaiko & Euzenat, 2005). Ces mé-

thodes visent à extraire les relations sémantiques (i.e. équivalence, subsomption, etc.) entre les entités (i.e. répertoires, catégories, concepts, propriétés etc.) issues de deux structures hiérarchiques telles que les système de fichiers, les schémas (de bases de données, de catalogues Web, etc.) ou encore les ontologies. La majorité des approches recensées s'appuient sur des mesures de similarité, et sont, par conséquence, pour la plupart restreintes à un alignement symétrique (relations d'équivalence).

Le modèle des règles d'association est fréquemment utilisé en fouille de données(Ceglar & Roddick, 2006). Les règles d'association sont des propositions de la forme "Si *prémisse* alors *conclusion*", notées $premisse \rightarrow conclusion$, qui représentent des tendances implicatives entre conjonctions d'attributs valués. Les règles d'association ont l'avantage d'être un modèle simple et intelligible pour représenter des connaissances explicites. De plus, cette technique d'apprentissage non-supervisée ne pré-requiert pas d'information particulière sur les connaissances à extraire, contrairement aux techniques d'apprentissage supervisées classiques tel que les arbres de décision ou prédicteurs bayésiens. Ces avantages ont motivés de nombreuses recherches avec, par exemple, la publication d'algorithmes d'extraction de règles (Agrawal *et al.*, 1993; Agrawal & Srikant, 1994) et la conception de mesures d'intérêt. Ces mesures d'intérêt permettent non seulement de vérifier la qualité de la tendance implicative d'une règle mais aussi de tester de nombreux autres aspects tels que la nouveauté, la significativité, la surprise, la non-trivialité, et l'applicabilité (Ceglar & Roddick, 2006).

A l'intersection de ces deux champs de recherche, nous avons proposé une méthode d'alignement extensionnelle et asymétrique de hiérarchies conceptuelles en utilisant le paradigme des règles d'association (David *et al.*, 2006). Notre approche repose fortement sur la nature asymétrique des règles d'association, permettant ainsi de découvrir des alignements implicatifs (relations de subsomption). Ce type d'alignement permet d'aider à caractériser plus précisément la nature des relations extraites (subsomption ou équivalence) par rapport à des méthodes uniquement basées sur des similarités. Contrairement à la majorité des approches d'alignement de schémas ou d'ontologies, notre méthode s'appuie fortement sur les données extensionelles fournies avec les structures. Ce type d'approche extensionelle est particulièrement adapté pour apparier des structures avec un schéma faiblement décrit (i.e. seulement des concepts organisés par une relation d'ordre partiel). Par exemple, notre approche peut fonctionner avec des taxonomies textuelles (répertoires web ou catalogues) ou avec des données semi-structurées. Notre méthode est divisée en deux étapes : (1) L'extraction et la sélection pour chaque concept d'un ensemble de termes significatifs ; (2) la découverte et sélection des implications génératrices entre les concepts.

Dans ce papier, nous proposons d'évaluer 6 mesures d'intérêt que nous avons sélectionnées selon des critères concernant leur sujet, leur nature et leur porté (Blanchard, 2005).

Dans un premier temps, nous présentons rapidement les principales approches d'alignement de hiérarchies de concepts. Ensuite, nous nous concentrons sur les mesures d'intérêt dans la fouille de règles d'association, et nous sélectionnons quelques indices d'après une classification basée sur leurs principales propriétés et leur sémantique. Dans la section suivante, nous exposons brièvement notre méthodologie d'alignement. Et finalement, nous procédons à la présentation et l'analyse des résultats expérimentaux.

Premièrement, nous évaluons la performance de notre méthode obtenue avec chacune des mesures sélectionnées. Ensuite, nous étudions les distributions des valeurs obtenues par ces mêmes mesures sur deux jeux de règles différents : un qui regroupe les alignements de référence et un autre qui regroupe un ensemble de relations non pertinentes.

2 Alignement de taxonomies textuelles

De nombreuses méthodes d'alignement de schémas de bases de données, d'ontologies, et de graphes ont été proposées dans la littérature. Elles utilisent des techniques d'apprentissage, de recherche d'information, ou encore de traitement du langage naturel. Ces disparités de points de vue rendent difficile la comparaison entre les différentes approches proposées dans la littérature. Cependant, quelques états de l'art sont disponibles : (Kalfoglou & Schorlemmer, 2003), (Rahm & Bernstein, 2001), (Shvaiko & Euzenat, 2005). En particulier, l'article de (Rahm & Bernstein, 2001) s'intéresse aux techniques d'appariement de schémas et propose une classification qui distingue les approches extensionnelles des approches intensionnelles. La majorité des travaux sur ce domaine concernent essentiellement des approches intensionnelles (une taxonomie de ces approches ainsi qu'une étude détaillée de leur caractéristiques est proposée par (Shvaiko & Euzenat, 2005)). Dans cet article, on s'intéressera seulement aux approches extensionnelles.

Le principe général des méthodes extensionnelles consiste à induire des relations, généralement d'équivalence, entre concepts ou catégories en comparant (généralement par une mesure de similarité) leur extensions textuelles. Comme les taxonomies ne sont, en principe, pas définies sur les mêmes bases de textes (instances), une première étape de pré-traitement est nécessaire dans le but de les rendre comparables. Selon le type de pré-traitement utilisé, nous dénotons deux grandes catégories d'approches : (1) celles qui s'appuient sur des méthodes de classification ; (2) celles qui utilisent des techniques issues de la recherche d'information et du traitement automatique du langage.

Les méthodes de la première famille utilisent pour la plupart des prédicteurs bayésien pour classer les textes de la première hiérarchie dans la deuxième et vice-versa. Cette étape de classification permet ainsi de représenter les concepts par un ensemble de documents partagés par les deux structures. Ensuite les concepts sont comparés en s'appuyant sur les documents partagés grâce à des mesures telles que la similarité de Jaccard (GLUE (Doan *et al.*, 2004)), la probabilité conditionnelle (oPLMap (Nottelmann & Straccia, 2005)). Elles peuvent également s'appuyer sur un test statistique comme la $\kappa - statistique$ utilisé dans Hichal (Ichise *et al.*, 2004).

La seconde approche consiste à représenter chaque concept par un ensemble termes caractéristiques extraits des documents. Le moyen le plus utilisé, issu de la recherche d'information, est de construire un vecteur pour chaque concept où les dimensions sont des termes et les composantes, les poids obtenus par une mesure telle que TF/IDF (utilisée dans SCM (Hoshiai *et al.*, 2004) et V-DOC (Qu *et al.*, 2006)) ou KullBack-Leiber (utilisée dans CAIMAN (Lacher & Groh, 2001)). Ensuite ces méthodes comparent des similarités entre concepts en s'appuyant sur la mesure de cosinus entre leur vecteurs caractéristiques. En fonction des méthodes, l'espace vectoriel peut prendre en compte les similarités entre dimensions (termes). Le second moyen, que nous utilisons, consiste

à représenter les concepts, non par des vecteurs, mais simplement par ensemble de termes préalablement extraits et sélectionnés en fonction de leur pertinence par rapport au concept. Ensuite la comparaison de deux concepts revient à comparer leur ensembles de termes respectifs.

3 Mesures d'intérêt pour l'évaluation de règles

Afin d'introduire et d'illustrer le modèle des règles d'association, nous nous appuyons sur un exemple typique d'application qui est l'étude du panier de la ménagère. Soit T ($card(T) = n$) l'ensemble des tickets de caisse (ou paniers) contenus dans la base de données. Soit A ($card(A) = n_a$) et B ($card(B) = n_b$) les ensembles de paniers contenant respectivement les produits a et b vendus dans le magasin. Une règle $a \rightarrow b$ sur T signifie que les consommateurs qui achètent le produit a ont également tendance à acheter le produit b. En principe, une telle règle sera d'autant meilleure qu'elle comportent beaucoup d'exemples (cad de paniers qui contiennent à la fois les produits a et b), noté $A \cap B$ ($card(A \cap B) = n_{ab}$) et relativement peu de contre-exemples (cad de paniers contenant le produit a mais pas le produit b), noté $A \cap \overline{B}$ ($card(A \cap \overline{B}) = n_{a\overline{b}}$).

Pour évaluer la qualité des règles d'association et aider les algorithmes d'extraction, deux mesures d'intérêt sont typiquement utilisées : Le support et la confiance. Le support ($s(a \rightarrow b) = n_{ab}/n$) représente la fréquence de la règle (i.e. le nombre de paniers qui contiennent à la fois les produits a et b par rapport au nombre de paniers contenus dans la base de données). La confiance ($c(a \rightarrow b) = n_{ab}/n_a$) évalue la qualité prédictive de la règle en mesurant la probabilité conditionnelle $P(b/a)$ (i.e. la probabilité qu'un panier contenant le produit a contienne également le produit b).

De nombreuses extensions du support et de la confiance ont été proposées dans la littérature ((Bayardo Jr. & Agrawal, 1999), (Tan *et al.*, 2004)) et on peut recenser plus de 40 mesures d'intérêt. Certains travaux se sont intéressés à définir les principes et les propriétés que doit respecter une bonne mesure d'intérêt ((Piatetsky-Shapiro, 1991), (Tan *et al.*, 2004), (Bayardo Jr. & Agrawal, 1999)).

Nous proposons de sélectionner parmi cet ensemble de mesures celles qui sont le mieux adaptées à notre objectif d'alignement. En autre, nous nous intéressons seulement aux mesures bornées (valeur maximum) afin de permettre à l'utilisateur de choisir plus facilement son seuil de sélection. D'autre part, en nous appuyant sur la classification proposée par (Blanchard *et al.*, 2005; Blanchard, 2005), nous proposons de sélectionner 6 mesures d'intérêts : la confiance, Loevinger, l'intensité d'implication (II), l'indice probabiliste d'écart à l'équilibre (Ipee), l'indice de la vraisemblence du lien (AVL), et l'indice de Jaccard.

La table 1 montre pour chaque mesure sélectionnée, sa portée (règle ($\rightarrow$), quasi-implication ($\Rightarrow$), quasi-conjonction ($\leftrightarrow$)), sa nature (statistique (S) ou descriptive (D)), son sujet (écart à l'indépendance (I) ou écart à l'équilibre (E)), la valeur fixe prise à l'indépendance ou à l'équilibre (fonction du sujet de la mesure) et sa définition.

Les critères utilisés dans la classification sont les suivants :

– *Le sujet*. Une mesure d'intérêt peut évaluer soit un écart à l'indépendance ou un écart à l'équilibre. L'indépendance est la situation dans laquelle le nombre de

Mesure	Portée	Nature	Sujet	Valeur fixe	Définition
II	$\Rightarrow$	S	I	0.5	$P(n_{a\bar{b}} < Poisson(\frac{n_a.n_{\bar{b}}}{n}))$
Loevinger	$\Rightarrow$	D	I	0	$1 - \frac{n.n_{a\bar{b}}}{n_a.n_{\bar{b}}}$
IPEE	$\rightarrow$	S	E	0.5	$P(n_{a\bar{b}} < Binomiale(n_a, 1/2))$
Confiance	$\rightarrow$	D	E	0.5	n_{ab}/n_b
AVL	$\leftrightarrow$	S	I	0.5	$P(n_{ab} > Poisson(\frac{n_a.n_b}{n}))$

TAB. 1 – Propriétés des mesures sélectionnées

contre-exemples de la règle est égal à celui attendu sous hypothèse d'indépendance statistique ($n_{a\bar{b}} = n_a.n_{\bar{b}}/n$). Les mesures qui évaluent un tel écart prennent une valeur fixe à l'indépendance. L'autre famille de mesures concerne celles qui évaluent l'écart par rapport à l'équilibre. La situation d'équilibre est atteinte quand le nombre de contre-exemples est égal à celui des exemples ($n_{a\bar{b}} = n_{ab}$). Ces mesures prennent une valeur fixe à l'équilibre.

– *La nature*. La nature d'une mesure d'intérêt peut être soit descriptive ou statistique. Les mesures descriptives ne sont pas influencées par une variation proportionnelle des cardinalités qu'elles prennent en compte. Ainsi une mesure descriptive m doit satisfaire l'égalité suivante : $m(n_a, n_b, n_{a\bar{b}}, n) = m(\alpha.n_a, \alpha.n_b, \alpha.n_{a\bar{b}}, \alpha.n)$ avec $\alpha > 0$. Inversement, les mesures statistiques varient avec la dilatation des cardinalités et ne respectent donc pas l'égalité ci-dessus. Selon les auteurs de cette classification, ce dernier type de mesures permet de confirmer statistiquement la validité des règles. Nous pouvons aussi ajouter qu'elles sont particulièrement adaptées pour détecter des règles originales. Par exemple, l'intensité d'implication décroît lorsque n_b croît, elle permet ainsi de privilégier les règles statistiquement valides qui ont une conclusion originale (en terme de support).
– *La portée*. La portée d'une mesure d'intérêt repose sur l'idée qu'une mesure évalue une proximité entre la règle et une configuration logique telle que l'implication, la conjonction ou encore l'équivalence (Blanchard, 2005). Pour qualifier la portée d'une mesure d'intérêt, nous utiliserons les termes de *quasi-implication*, *quasi-conjonction* et *quasi-équivalence* car les règles ne sont pas des propositions logiques strictes vu qu'elles peuvent avoir des contre-exemples. De plus, certaines mesures évaluent seulement la tendance que la conclusion de la règle soit vérifiée lorsque la prémisse l'est. De telles mesures comme la confiance ou IPEE (Blanchard, 2005) ne sont pas associées à une configuration logique et portent ainsi seulement sur l'évaluation de la règle.

4 Méthode d'alignement extensionnelle et asymétrique

Notre méthode (figure 1) prend en entrée deux hiérarchies de concepts organisés par une relation d'ordre partiel et connectés à un ensemble de documents textuels.

Nous définissons une hiérarchie conceptuelle $\mathcal{H}$ par un quadruplet $\mathcal{H} = (C, \leq, D, \sigma_0)$. C représente l'ensemble des concepts et D, l'ensemble des documents. $\leq$ est la rela-

tion d'ordre organisant les concepts en taxonomie, et σ_0 est une relation qui associe un ensemble de documents à chaque concept (pour $c \in C$, $\sigma_0(c)$ est l'ensemble des documents associés à c). A partir de $\leq$, nous étendons la relation σ_0 en $\sigma : \sigma(c) = \bigcup_{c' \leq c} \sigma_0(c')$.

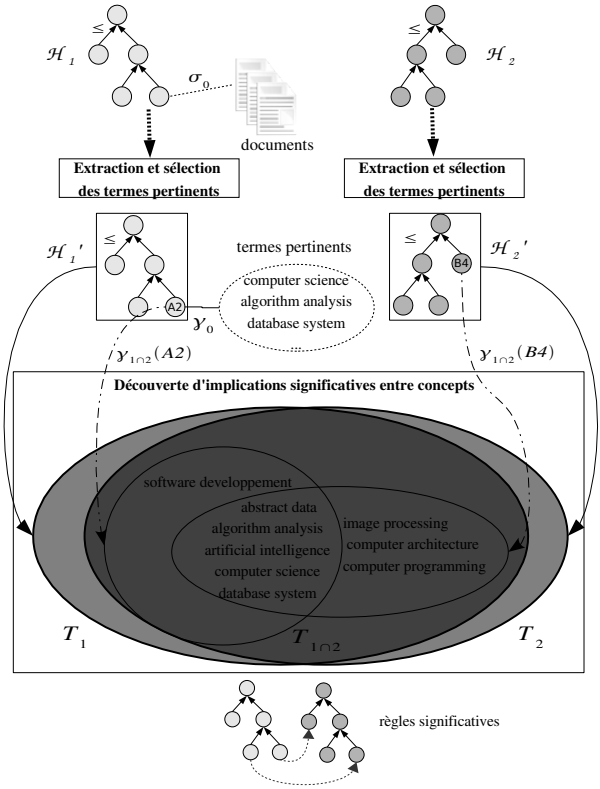

FIG. 1 – Schéma de la méthode

La première phase de notre méthode permet de transformer une hiérarchie $\mathcal{H}$ définie sur des documents, en une hiérarchie $\mathcal{H}' = (C, \leq, T, \gamma_0)$, définie sur des termes (T est l'ensemble des termes pertinents sélectionnés et γ_0 représente la relation qui associe un ensemble de termes pertinents pour chaque concept). Cette première phase permet tout d'abord de représenter chaque texte par un ensemble de termes binaires. Cette indexation terminologique est réalisée au moyen du logiciel ACABIT Daille (2003) sur un texte préalablement étiqueté grammaticalement. Nous avons choisi d'utiliser les termes binaires car ils ne sont pas ambigus contrairement aux termes simples. Ensuite, nous procédons l'évaluation des règles d'association $terme \rightarrow concept$ par l'intensité d'implication. Un terme t sera considéré comme pertinent pour le concept c (et donc sélectionné) si la règle $t \rightarrow c$ a une valeur d'intensité d'implication $\varphi(t \rightarrow c)$ supérieure au seuil de sélection fixé φ_t. Intuitivement, un terme t sera pertinent pour le concept c si t tend à apparaitre seulement dans des documents associés au concept c. Ainsi,

$\gamma(c) = \{t|\varphi(t \rightarrow c) \geq \varphi_t\}$. Finalement, à partir de $\leq$, nous étendons la relation γ_0 en $\gamma : \gamma(c) = \bigcup_{c' \leq_c} \gamma_0(c')$ afin d'obtenir un morphisme de $(C, \leq)$ dans $(2^T, \subseteq)$.

A partir de deux hiérarchies $\mathcal{H}'_1 = (C_1, \leq_1, T_1, \gamma_1)$ et $\mathcal{H}'_2 = (C_2, \leq_2, T_2, \gamma_2)$, la deuxième phase de notre approche permet d'extraire l'ensemble de règles d'association significatives de la forme $a \rightarrow b$ où $a \in C_1$ et $b \in C_2$. Pour cela, nous choisissons de travailler sur l'ensemble des termes pertinents communs aux deux hiérarchies noté $T_{1\cap 2} = T_1 \cap T_2$. Nous fusionnons alors les relations d'association des termes aux concepts pour obtenir la relation $\gamma_{1\cap 2}(c)$.

Lors de la phase de sélection des règles d'association, de nombreuses règles redondantes peuvent être générées. Ainsi, nous proposons deux critères permettant de définir la significativité d'une règle entre concepts.

Une règle $a \rightarrow b$ (avec $a \in C_1$ et $b \in C_2$) sera dite "significative" si : (1) $\varphi(a \rightarrow b) \leq \varphi_r$ et (2) $\forall x \geq a,\ \forall y \leq b,\ \varphi(x \rightarrow y) \leq \varphi(a \rightarrow b)$

Le premier critère permet de garantir la qualité de la quasi-implication entre concepts. Le deuxième critère permet de réduire la redondance, car il ne sélectionne que des règles n'ayant pas de règles génératrices avec une valeur d'intensité supérieure. Une règle $a \rightarrow b$ est dite génératrice si il n'existe pas de règle $x \rightarrow y$ ($x \leq a$, $b \leq y$, et $x \rightarrow y \neq a \rightarrow b$) ayant $\varphi(a \rightarrow b) \leq \varphi(x \rightarrow y)$.

5 Résultats expérimentaux

Les expérimentations présentées dans cette section ne concernent que la phase de sélection des règles. Dans un premier temps, nous comparons l'efficacité des mesures grâce à la f-mesure qui agrège la précision et le rappel relativement à un alignement de référence. Ensuite, nous proposons une analyse des distributions des valeurs prises par les différentes mesures sur deux alignements manuels : $R+$, l'alignement positif constitué de l'alignement de référence fourni avec le jeux de test ; $R-$, l'alignement négatif contenant des relations totalement incohérentes. Les deux ensembles $R+$ et $R-$ sont de même cardinalité.

5.1 Données analysées

Le jeu de test utilisé a été proposé par Doan *et al.* (2004). Il est composé de deux catalogues de cours proposés par les universités de Cornell et Washington. Les descriptions textuelles des cours sont organisées de manière hiérarchique en école et collèges et ensuite en départements et centres. Les deux hiérarchies contiennent respectivement 166 et 176 concepts auxquels sont associés 4360 et 6957 descriptions de cours. Le jeu de test est également fourni avec un ensemble de 54 relations d'alignement entre les concepts du catalogue Cornell et le catalogue Washington. Cet ensemble de relations de référence n'est cependant constitué que d'équivalences.

5.2 Evaluation des mesures d'intérêt

Cette section présente les résultats obtenus par notre méthode d'alignement testée avec chacune des mesures sélectionnées. Pour chaque mesure, nous avons fait varier le

seuil de sélection des règles (φ_r) entre 0 et 1 avec un pas de 1%. Pour chaque valeur de seuil, l'algorithme est exécuté deux fois : tout d'abord pour extraire les implications entre Cornell vers Washington puis de Washington vers Cornell. A partir des ensembles d'implications extraites, nous ne gardons que les relations d'équivalence déduites en suivant la règle "<si $a \rightarrow b$ et $b \rightarrow a$, alors $a \leftrightarrow b$">.

Dans le but d'évaluer la pertinence des résultats vis-à-vis de l'alignement de référence, nous utilisons deux métriques classiques issues de la recherche d'information : la précision (P) et le rappel (R). Ces mesures sont définies ainsi : soit F l'ensemble des relations trouvées par notre approche et R l'ensemble des relations de référence. La précision ($precision = card(F \cap R)/card(F)$) mesure le ratio entre nombre de bonnes relations (i.e. les relations qui sont à la fois dans le résultat et l'alignement de référence) et le nombre de relations trouvées par notre approche. Le rappel ($rappel = card(F \cap R)/card(R)$) mesure le ratio entre le nombre de bonnes relations et le nombre de relations contenues dans l'alignement de référence. Finalement, ces deux mesures sont agrégées par la f-mesure (similarité de Dice entre les ensembles F et R) : $f - mesure = 2.precision.rappel/precision + rappel$.

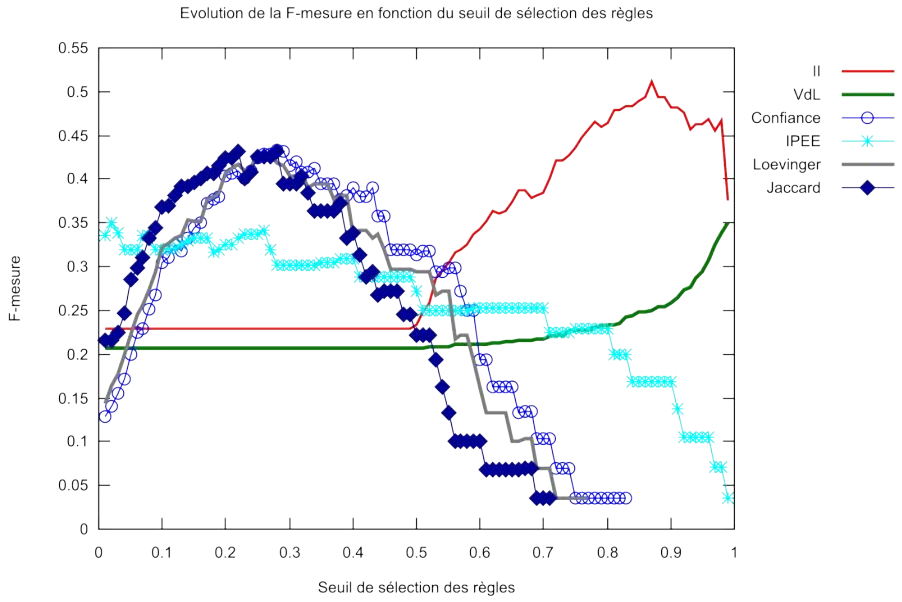

FIG. 2 – Evolution de la f-mesure

La figure 2 montre que les 3 mesures descriptives, Loevinger, la confiance et Jaccard, sont proches. La valeur maximale de f-mesure est atteinte pour un seuil φ_r d'environ $0,3$. Cette valeur seuil reste cohérente avec la sémantique de l'indice de Loevinger puisqu'elle dépasse le seuil d'indépendance (qui est de 0). Par contre, cela n'est pas le cas pour les mesures de confiance et de Jaccard pour lesquelles la situation d'équilibre,

qui se situe à $0,5$, n'est pas atteinte. L'évolution de la valeur de la f-mesure avec Ipee est décroissante en fonction du seuil φ_r croissant. Cependant, il est plus robuste que Loevinger, la confiance et Jaccard face à l'augmentation du seuil de sélection φ_r. L'indice de vraisemblance du lien (AVL) obtient une courbe de f-mesure croissante mais son maximum est faible. Finalement, les meilleurs résultats sont obtenus avec l'intensité d'implication. Son meilleur score en terme de f-mesure est de $0,51$ pour un seuil de sélection fixé à $0,9$. Vis-à-vis de la sémantique de l'indice, l'évolution de la f-mesure est parfaitement cohérente : elle reste stable jusqu'à la situation d'indépendance (seuil de $0,5$) avant d'augmenter.

5.3 Distributions des mesures d'intérêt

Avec cette expérimentation, nous nous proposons de comparer comment les mesures évaluent les relations d'alignement indépendamment de notre algorithme de sélection des règles. Nous utilisons cependant la phase de sélection des termes significatifs en utilisant l'intensité d'implication avec un seuil de $0,9$. Nous dressons les distributions des valeurs prises par les différentes mesures sur les deux alignements $R+$ et $R-$

Pour le premier test, chaque relation est représentée par un couple (A, B). Pour chacune de ces relations (dans le cas des mesures asymétriques), nous avons évalué les règles $A \rightarrow B$ et $B \rightarrow A$. Pour chaque couple, nous retenons la meilleure valeur.

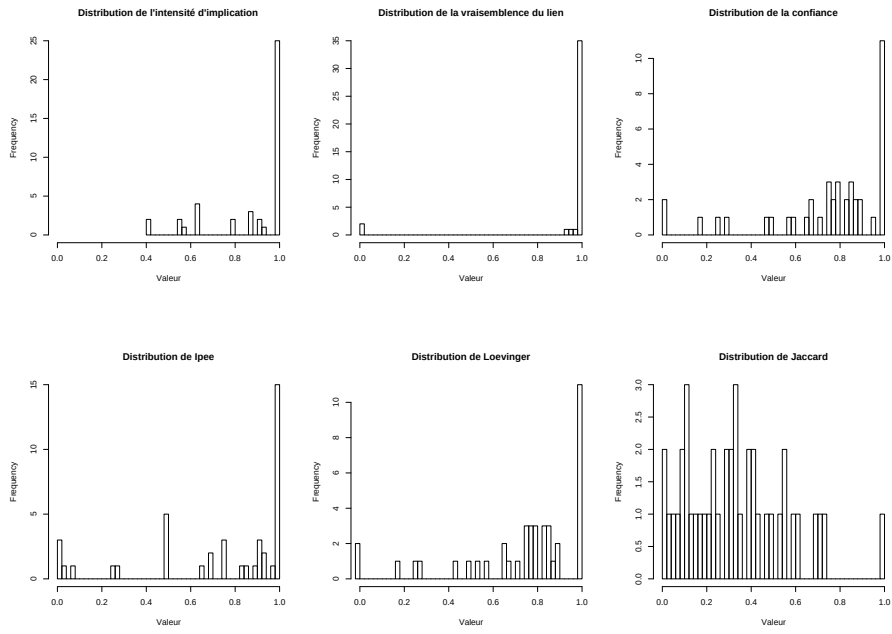

FIG. 3 – Distributions des valeurs prises par les mesures sur l'ensemble de relations pertinentes $R+$

Les distributions des valeurs de la confiance et de Ipee (figure 3) montre qu'une majorité des règles sont bien évaluées en terme d'équilibre (valeurs au dessus de 0.5). Dans les deux cas 6 ou 7 règles sont en dessous du seuil. Cependant, la mesure de Jaccard évalue beaucoup de règles en dessous du seuil d'équilibre. D'un autre coté, la mesure de Loevinger montre que les relations sont pertinentes en termes d'écart à l'indépendance vu que la grande majorité des règles ont des valeurs supérieures à 0 (2 relations seulement sont évaluées en dessous de 0). Les deux mesures probabilistes d'écart à l'indépendance, l'intensité d'implication et la vraisemblance du lien évaluent la majorité des relations avec de bonnes valeurs (deux relations seulement sont évaluées en dessous du seuil d'indépendance de $0,5$). L'intensité d'implication a cependant tendance à plus distribuer les règles dans l'intervalle $[0,5-1]$.

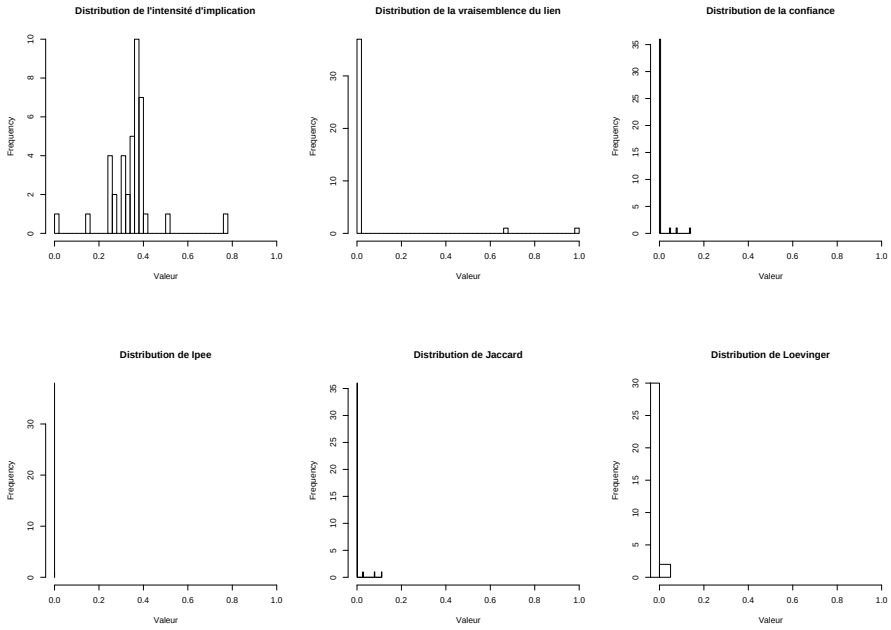

FIG. 4 – Distributions des valeurs prises par les mesures sur l'ensemble de relations non pertinentes $R-$

Pour cette deuxième série de distributions, nous avons évalué un ensemble de relations non pertinentes. Dans ce cas, nous avons retenu pour chaque couple de concepts la valeur minimale obtenue sur l'évaluation des deux règles qui en sont issues.

Sur la figure 4, les valeurs des indices d'écart à l'équilibre (confiance, Jaccard et Ipee) sont toutes proches de 0. Loevinger évalue 30 règles en dessous 0, ce qui est accord avec la sémantique de l'indice. L'intensité d'implication et la vraisemblance du lien sont aussi adaptées dans ce cas puisque seulement deux règles sont évaluées au dessus de $0,5$. L'intensité d'implication a encore tendance à répartir les règles entre 0

et $0,5$ contrairement aux autres indices. Les 6 mesures fonctionnent bien dans le cas d'un alignement composé de relations incohérentes.

En comparant les deux ensembles de distributions, les mesures d'écart à l'indépendance ont une meilleure capacité à discriminer les règles pertinentes des mauvaises. Nous pouvons aussi remarquer que dans ce contexte, le nombre de contre-exemples nécessaire pour atteindre la situation d'équilibre est moins élevé que celui nécessaire pour atteindre la situation d'indépendance.

6 Conclusion

Nous avons étudié l'utilisation de différentes mesures dans une approche d'alignement de hiérarchies textuelles basée sur le modèle des règles d'association. Cette méthode d'alignement extensionnelle a l'originalité d'utiliser l'aspect asymétrique des règles d'association dans le but de découvrir des relations de subsomption entre entités issues de deux hiérarchies. Dans ce cadre, nous avons réutilisé les travaux portant sur les mesures d'intérêts utilisées pour l'évaluation des règles d'association et nous avons sélectionné un échantillon de mesures selon différents critères (sujet, nature et portée). Nous avons mené deux évaluations différentes des mesures sélectionnées sur un jeu de test portant sur un alignement de catalogues. Les deux expérimentations montrent que les mesures d'écart à l'indépendance sont plus adaptées à l'évaluation de telles règles. Dans ce contexte où les phénomènes étudiés sont petits, la situation d'équilibre est atteinte plus rapidement que celle d'indépendance lors de l'augmentation du nombre de contre-exemples. Les résultats montrent par ailleurs que les mesures descriptives (confiance, Loevinger, Jaccard) ont des comportements très similaires. Finalement, les meilleurs résultats ont été obtenus avec l'intensité d'implication, une mesure probabiliste d'écart à l'indépendance dont la portée est la quasi-implication. Cette étude s'est limitée à l'étude de l'influence des mesures utilisées sur notre approche. Il serait également intéressant de mener de telles évaluations des mesures sur d'autres méthodes extensionnelles telles que celle présentées dans la section d'état de l'art.

Références

AGRAWAL R., IMIELINSKI T. & SWAMI A. (1993). Mining association rules between sets of items in large databases. In *the proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data*, p. 207–216 : ACM Press.

AGRAWAL R. & SRIKANT R. (1994). Fast algorithms for mining association rules. In J. BOCCA, M. JARKE & C. ZANIOLO, Eds., *Proc. 20th Int. Conf. Very Large DataBases*, p. 487–499 : Morgan Kaufmann.

BAYARDO JR. R. J. & AGRAWAL R. (1999). Mining the most interestingness rules. *Proc. of the 5th ACM SIGKDD Int. Conf. On Knowledge Discovery and Data Mining*, p. 145–154.

BLANCHARD J. (2005). *A visualization system for interactive mining, assessment, and exploration of association rules*. PhD thesis, University of Nantes.

BLANCHARD J., GUILLET F., GRAS R. & BRIAND H. (2005). Using information-theoretic measures to assess association rule interestingness. In *the proc. of the 15th IEEE Int. Conf. on Data Mining ICDM'05*, p. 66–73 : IEEE Computer Society.

CEGLAR A. & RODDICK J. F. (2006). Association mining. *ACM Comput. Surv.*, **38**(2), 5.

DAILLE B. (2003). Conceptual structuring through term variations. In F. BOND, A. KORHONEN, D. MACCARTHY & A. VILLACICENCIO, Eds., *ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16.

DAVID J., GUILLET F., GRAS R. & BRIAND H. (2006). Conceptual hierarchies matching : an approach based on discovery of implication rules between concepts. In *the proc. of the 17th European Conf. on A.I.*, p. pages 357–361.

DOAN A., MADHAVAN J., DOMINGOS P. & HALEVY A. (2004). Ontology matching : a machine learning approach. In S. STAAB & R. STUDER, Eds., *Handbook on Ontologies in Information Systems*, p. 397–416. Springer-Velag.

HOSHIAI T., YAMANE Y., NAKAMURA D. & TSUDA H. (2004). A semantic category matching approach to ontologies alignment. In *the 3rd int. workshop on Eval. of Ontology Based Tools*.

ICHISE R., M.HAMASAKI & TAKEDA H. (2004). Discovering relationships among catalogs. In E. SUZUKI & S. ARIKAWA, Eds., *7th Int. Conf. Discovery Science 2004*, volume 3245 of *LNCS*, p. 371–379 : Springer.

KALFOGLOU Y. & SCHORLEMMER M. (2003). Ontology mapping : the state of the art. *Knowledge Engineering Review*, **18**(1), 1–31.

LACHER M. S. & GROH G. (2001). Facilitating the exchange of explicit knowledge through ontology mappings. In *the proc. of 14th Int. Florida A.I. Research Society Conf.*, p. 305–309 : AAAI Press.

NOTTELMANN H. & STRACCIA U. (2005). A probabilistic, logic-based framework for automated web directory alignment. In Z. MA, Ed., *Soft Computing in Ontologies and the Semantic Web*, Studies in Fuzziness and Soft Computing. Springer Verlag.

PIATETSKY-SHAPIRO G. (1991). Discovery, analysis and presentation of strong rules. In G. PIATETSKY-SHAPIRO & W. FRAWLEY, Eds., *Knowledge Discovery in Databases*, p. 229–248.

QU Y., HU W. & CHENG G. (2006). Constructing virtual documents for ontology matching. In *the proc. of the 15th int. conf. on WWW*, p. 23–31, New York, NY, USA : ACM Press.

RAHM E. & BERNSTEIN P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, **10**(4), 334–350.

SHVAIKO P. & EUZENAT J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics IV*, **4**(LNCS 3730), 146–171.

TAN P.-N., KUMAR V. & SRIVASTAVA J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, **29(4)**, 293–313.